Structural counterfactuals: A brief introduction

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
Phone: (310) 825-3243
Fax: (310) 794-5057
E-mail: judea@cs.ucla.edu

Abstract

Recent advances in causal reasoning have given rise to a computational model that emulates the process by which humans generate, evaluate and distinguish counterfactual sentences. Contrasted with the "possible worlds" account of counterfactuals, this "structural" model enjoys the advantages of representational economy, algorithmic simplicity and conceptual clarity. This introduction traces the emergence of the structural model and gives a panoramic view of several applications where counterfactual reasoning has benefited problem areas in the empirical sciences.

1 Introduction

One of the most striking phenomena in the study of conditionals is the ease and uniformity with which people evaluate counterfactuals. To witness, the majority of people would accept the statement: S_1: "If Oswald didn't kill Kennedy, someone else did," but few, if any, would accept its subjunctive version: S_2: "If Oswald hadn't killed Kennedy, someone else would have." For students of conditionals, these canonical examples (attributed to Ernst Adams, (1975)) represent a compelling proof of the ubiquity of the indicative/subjunctive distinction, and of the amazing capacity of humans to process, evaluate and form consensus about counterfactuals.

Yet, not many students of conditionals asked the next question: How do we, humans, reach such consensus? More concretely, what mental representation permits such consensus to emerge from the little knowledge we have about Oswald, Kennedy and 1960's Texas, and what algorithms would need to be postulated to account for the swiftness, comfort and confidence with which such judgments are issued.

The basic thesis of structural counterfactuals (Balke & Pearl, 1995; Pearl, 2000) is that counterfactuals are generated and evaluated by symbolic operations on a model that represents an agent's beliefs about functional relationships in the world. The procedure can be viewed as a concrete implementation of Ramsey's idea (Ramsey, 1929), according

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Structural counterfactuals: A brief introduction	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California, Los Angeles,Computer Science Department,Los Angeles,CA,92295-1596**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Recent advances in causal reasoning have given rise to a computational model that emulates the process by which humans generate, evaluate and distinguish counter- factual sentences. Contrasted with the possible worlds" account of counterfactuals this structural" model enjoys the advantages of representational economy, algorithmic simplicity and conceptual clarity. This introduction traces the emergence of the struc- tural model and gives a panoramic view of several applications where counterfactual reasoning has bene ted problem areas in the empirical sciences.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**10**	

to which a conditional is accepted if the consequent is true after we add the antecedent (hypothetically) to our stock of beliefs and make whatever minimal adjustments are required to maintain consistency (Arlo-Costa, 2009). In the indicative case, we simply add the antecedent A as if we received a new evidence that affirms its truth and discredits whatever previous evidence we had for its negation. In the subjunctive case, we establish the truth of A by changing the model itself.

Taking Kennedy's assassination as a working example, the distinction is as follows:

To evaluate the indicative conditional S_1 ("If Oswald didn't kill Kennedy, someone else did") we start by assigning truth values to variables that are known (or believed) to be true in the story. In our case, we start with the common knowledge that Kennedy was in fact killed, so, adding the hypothetical fact that Oswald did not kill Kennedy, it must be that someone-else killed him.

The evaluation of the subjunctive conditional S_2 ("If Oswald hadn't killed Kennedy, someone else would have") demands a different procedure. S_2 calls for rolling back history as we know it, and rerun it under different conditions where, for some unknown reason, Oswald refrains from shooting Kennedy. The key difference between the two procedures lies in holding Kennedy's death true in the indicative case but leaving it uncommitted in the subjunctive case.

In Section 2 of this paper, I will present simple algorithms that reliably interpret subjunctive conditionals, and cast these algorithms in the context of the general theory of structural counterfactuals. I will briefly compare the structural account of counterfactuals to the "possible worlds" account of Lewis (1973) and defend my preference of the former. In Section 3, I will demonstrate how this model has given rise to an effective methodology of causal inference in several of the empirical sciences, and how it has helped resolve practical questions, from policy evaluation to mediation analysis to generalizing conclusions across experimental studies.

2 An outline of the structural theory

The distinctions illustrated in the preceding section are part of a general theory of counterfactuals that I named "structural" (Pearl, 2000, Chapter 7) in honor of its origin in the structural equation models developed by econometricians in the 1940-50's (Haavelmo, 1943; Simon, 1953; Hurwicz, 1950; Marschak, 1953).

At the center of the theory lies a "structural model," M, consisting of two sets of variables, U and V, and a set F of functions that determine how values are assigned to each variable $V_i \in V$. Thus for example, the equation

$$v_i = f_i(v, u)$$

describes a physical process by which Nature *examines* the current values, v and u, of all variables in V and U and, accordingly, *assigns* variable V_i the value $v_i = f_i(v, u)$. The variables in U are considered "exogenous," namely, background conditions for which no explanatory mechanism is encoded in model M. Every instantiation $U = u$ of the exogenous variables uniquely determines the values of all variables in V and, hence, if we assign a probability $P(u)$ to U, it defines a probability function $P(v)$ on V.

The basic counterfactual entity in structural models is the sentence: "Y would be y had X been x in situation $U = u$," denoted $Y_x(u) = y$. The key to interpreting counterfactuals is to treat the subjunctive phrase "had X been x" as an instruction to make a "minimal" modification in the current model, so as to ensure the antecedent condition $X = x$. Such a minimal modification amounts to replacing the equation for X by a constant x. This replacement permits the constant x to differ from the actual value of X (namely $f_X(v, u)$) without rendering the system of equations inconsistent, thus allowing all variables, exogenous as well as endogenous, to serve as antecedents.

Letting M_x stand for a modified version of M, with the equation(s) of X replaced by $X = x$, the formal definition of the counterfactual $Y_x(u)$ reads:

$$Y_x(u) \triangleq Y_{M_x}(u). \tag{1}$$

In words: The counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the "surgically modified" submodel M_x.[1] Galles and Pearl (1998) and Halpern (1998) have given a complete axiomatization of structural counterfactuals, embracing both recursive and non-recursive models (see also Pearl, 2009a, Chapter 7).

Since the distribution $P(u)$ induces a well defined probability on the counterfactual event $Y_x = y$, it also defines a joint distribution on all Boolean combinations of such events, for instance '$Y_x = y$ AND $Z_{x'} = z$,' which may appear contradictory, if $x \neq x'$. For example, to answer retrospective questions, such as whether Y would be y_1 if X were x_1, given that in fact Y is y_0 and X is x_0, we need to compute the conditional probability $P(Y_{x_1} = y_1|Y = y_0, X = x_0)$ which is well defined once we know the forms of the structural equations and the distribution of the exogenous variables in the model.

In general, the probability of the counterfactual sentence $P(Y_x = y|e)$, where e is any propositional evidence, can be computed by the 3-step process (illustrated in Pearl, 2009a, p. 207);

Step 1 (***abduction***): Update the probability $P(u)$ to obtain $P(u|e)$.

Step 2 (***action***): Replace the equations corresponding to variables in set X by the equations $X = x$.

Step 3 (***prediction***): Use the modified model to compute the probability of $Y = y$.

In temporal metaphors, Step 1 explains the past (U) in light of the current evidence e; Step 2 bends the course of history (minimally) to comply with the hypothetical antecedent $X = x$; finally, Step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$.

It can be shown (Pearl, 2000, p. 76) that this procedure can be given an interpretation in terms of "imaging" (Lewis, 1973) – a process of "mass-shifting" among possible worlds – provided that (a) worlds with equal histories should be considered equally similar and (b) equally-similar worlds should receive mass in proportion to their prior probabilities (Joyce,

[1]Simon and Rescher (1966) came close to this definition but, lacking the "wiping out" operator, could not reconcile the contradiction that ensues when an observation $X = x'$ clashes with the antecedent $X = x$ of the counterfactual Y_x.

2009; Pearl, 2000, pp. 76; 2010). Because "similarities" are thus shaped by causal-temporal priorities, the structural account does not suffer from classical paradoxes that plague "similarity by appearance" (Taylor & Dennett, 2011). For example, the sentence "Had Nixon pressed the button, a nuclear war would have started" is accepted as true, despite Fine's ((1975)) "more similar" scenario in which someone had disconnected the switch. Fine's scenario is not minimally sufficient to ensure the antecedent "pressed the button."

In (Pearl, 2000, p. 239), I remarked: "In contrast with Lewis's theory, [structural] counterfactuals are not based on an abstract notion of similarity among hypothetical worlds; instead they rest directly on the mechanisms (or 'laws,' to be fancy) that govern those worlds and on the invariant properties of those mechanisms. Lewis's elusive 'miracles' are replaced by principled mini-surgeries, $do(X = x)$, which represent a minimal change (to a model) necessary for establishing the antecedent $X = x$ (for all u). Thus, similarities and priorities—if they are ever needed—may be read into the $do(\cdot)$ operator as an afterthought (see (Pearl, 2000, Eq. (3.11)) and (Goldszmidt & Pearl, 1992)), but they are not basic to the analysis."

2.1 The two principles of causal inference

Before describing specific applications of the structural theory, it will be useful to summarize its implications in the form of two "principles." The entire set of tools needed for solving causal and counterfactuals problems are based on only these two:

Principle 1: "The law of structural counterfactuals."

Principle 2: "The law of structural independence."

The first principle is described in Eq. (1) and instructs us how to compute counterfactuals from a structural model. It thus allows us to define formally which counterfactual is true in a given model M and in any given circumstance ($U = u$), and to express communicable assumptions about reality in terms of counterfactual sentences. Likewise, when probabilities are defined on U, principle 1 permits us to compute probabilities of a counterfactual, to determine if one counterfactual depends on another given a third and, most importantly, to determine what assumptions one must make about reality in order to infer probabilities of counterfactuals from either experimental or passive observations.

Principle 2 instructs us how to detect conditional independencies in the data from the structure of the model, that is, from the graph that describes the functional relationships between the variables. Remarkably, regardless of the functional form of the equations in the model and regardless of the distribution of the exogenous variables U, if the disturbances are mutually independent and the model is recursive, the distribution $P(v)$ of the endogenous variables must obey certain conditional independence relations, stated roughly as follows: whenever sets X and Y are "separated" by a set Z in the graph, X is independent of Y given Z in the probability.[2]

[2]The "separation" criterion requires that all paths between X and Y be intercepted by Z, with special handling of paths containing head-to-head arrows (Pearl, 2000, pp. 16–18). In linear models, Principle 2 is valid for non-recursive models as well.

This powerful theorem, called d-separation (Pearl, 2000, pp. 16–18) constitutes the link between causal assumptions encoded in the model and the observed data. It serves as the basis for causal discovery algorithms (Pearl & Verma, 1991; Spirtes, Glymour, & Scheines, 1993) as well as deciding identification and testing model misspecification.

3 Summary of applications

Since its inception (Balke & Pearl, 1995) this counterfactual model has provided mathematical solutions to a vast number of lingering problems in policy analysis and retrospective reasoning. In the context of decision making, for example, a rational agent is instructed to maximize the expected utility

$$EU(x) = \sum_y P(Y_x = y)U(y) \tag{2}$$

over all options x. Here, $U(y)$ stands for the utility of outcome $Y = y$ and $P(Y_x = y)$ stands for the probability that outcome $Y = y$ would prevail, had action $do(X = x)$ been performed and condition $X = x$ firmly established.[3]

The central question in many of the empirical sciences is that of *identification*: Can we predict the effect of a contemplated action $do(X = x)$ or, in other words, can the post-intervention distribution, $P(Y_x = y)$, be estimated from data generated by the pre-intervention distribution, $P(z, x, y)$? Clearly, since the prospective counterfactual Y_x is generally not observed, the answer must depend on the agent's model M and then the question reduces to: Can $P(Y_x = y)$ be estimated from a combination of $P(z, x, y)$ and a graph G that encodes the structure of M.

This problem has been solved by deriving a precise characterization of what Skyrms (1980) called "KD-partition," namely, a set S of observed variables that permits $P(Y_x = y)$ to be written in terms of Bayes conditioning or, "adjusting for" S:

$$P(Y_x = y) = \sum_s P(y|x, s)P(s).$$

The solution came to be known as the back-door criterion (Pearl, 1995), stating (roughly) that a set S of variables is admissible for adjustment if it "blocks" every path between X and Y that ends with an arrow into X. Hagmayer and Sloman (2009) provide some evidence that this is exactly what people do. Tian and Pearl (2002) and Shpitser and Pearl (2007) further expanded this result and established a criterion that permits (or forbids) the assessment of $P(Y_x = y|Z = z)$ by any method whatsoever, including the use of X-dependent covariates Z (Pearl, 2009a, pp. 339–341), and time-varying sets of antecedent variables X.

Prospective counterfactual expressions of the type $P(Y_x = y)$ are concerned with predicting the average effect of hypothetical actions and policies and can, in principle,

[3]Equation (2) represents the dictates of Causal Decision Theory (CDT) Stalnaker (1981); Lewis (1973); Gardenfors (1988) and Joyce (1999) – the pitfalls of Evidential Decision Theory are well documented (see (Skyrms, 1980; Pearl, 2000, pp. 108–9)), and need not be considered.

be assessed from experimental studies in which X is randomized. Retrospective counterfactuals, on the other hand, like S_2 in the Oswald scenario, consist of variables at different hypothetical worlds (different subscripts) and these may or may not be testable experimentally. In epidemiology, for example, the expression $P(Y_{x'} = y'|x, y)$ may stand for the fraction of patients who recovered (y) under treatment (x) who would not have recovered (y') had they not been treated (x'). This fraction cannot be assessed in experimental study, for the simple reason that we cannot re-test patients twice, with and without treatment. A different question is therefore posed: which counterfactuals can be tested, be it in experimental or observational studies. This question has been given a mathematical solution in (Shpitser & Pearl, 2007). It has been shown, for example, that in linear systems, $E(Y_x|e)$ is estimable from experimental studies whenever the prospective effect $E(Y_x)$ is estimable in such studies. Likewise, the counterfactual probability $P(Y_{x'}|x)$, also known as the effect of treatment on the treated (ETT) is estimable from observational studies whenever an admissible S exists for $P(Y_x = y)$ (Shpitser & Pearl, 2009).

Retrospective counterfactuals have also been indispensable in conceptualizing direct and indirect effects (Baron & Kenny, 1986; Robins & Greenland, 1992; Pearl, 2001), which require nested counterfactuals in their definitions. For example, to evaluate the direct effect of treatment $X = x'$ on individual u, un-mediated by a set Z of intermediate variables, we need to construct the nested counterfactual $Y_{x',Z_x(u)}$ where Y is the effect of interest, and $Z_x(u)$ stands for whatever values the intermediate variables Z would take had treatment not been given.[4] Likewise, the average *indirect effect*, of a transition from x to x' is defined as the expected change in Y affected by holding X constant, at $X = x$, and changing Z, hypothetically, to whatever value it would have attained had X been set to $X = x'$.

This counterfactual formulation has enabled researchers to derive conditions under which direct and indirect effects are estimable from empirical data (Pearl, 2001; Petersen, Sinisi, & van der Laan, 2006) and to answer such questions as: "Can data prove an employer guilty of hiring discrimination?" or, phrased counterfactually, "what fraction of employees owes its hiring to sex discrimination?"

These tasks are performed using a general estimator, called the Mediation Formula (Pearl, 2001, 2009b, 2012a), which is applicable to nonlinear models with discrete or continuous variables, and permits the evaluation of path-specific effects with minimal assumptions regarding the data-generating process (Pearl, 2012b, 2012c).

Finally, as the last application, I point to a recent theory of "transportability" (Pearl & Bareinboim, 2011) which provides a formal solution to the century-old problem of "external validity" (Campbell & Stanley, 1966); i.e., under what conditions can experimental findings be transported to another environment, how the results should be calibrated to account for environmental differences, and what measurements need be taken in each of the two environments to license the transport.

The impact of the structural theory in the empirical sciences does not prove, of course, its merits as a cognitive theory of counterfactual reasoning. The evidence is in fact mixed on this issue (see Sloman and Lagnado (2005), versus Rips, this issue. Also see Kaufmann,

[4]Note that conditioning on the intermediate variables in Z would generally yield the wrong answer, due to unobserved "confounders" affecting both Z and Y. Moreover, in non linear systems, the value at which we hold Z constant will affect the result (Pearl, 2000, pp. 126-132).

this issue). It proves nevertheless that in the arena of policy evaluation and decision making the theory is compatible with investigators states of belief and, whenever testable, its conclusions have withstood the test of fire.

4 Conclusions

This introduction started with the enigma of consensus: "What mental representation permits such consensus to emerge from the little knowledge we have about Oswald, Kennedy and 1960's Texas, and what algorithms would need to be postulated to account for the swiftness, comfort and confidence with which such judgments are issued." The very fact that people communicate with counterfactuals already suggests that they share a similarity measure, that this measure is encoded parsimoniously in the mind, and hence that it must be highly structured.

The theory of structural counterfactuals offers a solution to the consensus enigma. It presents conceptually clear and parsimonious encoding of knowledge from which causes, counterfactuals, and probabilities of counterfactuals can be derived by effective algorithms. It further carries the potential of teaching robots to communicate in the language of counterfactuals and eventually acquire an understanding of notions such as responsibility and regret, pride and free will.

The theory has given rise to major breakthroughs in the methodology of the empirical sciences.

Acknowledgments

I am grateful to Daniel Dennett and Christopher Taylor for discussing the structural theory of counterfactuals and thus enticing me to write this exposition, and to the editor, Steven Sloman, for giving me the opportunity to present these ideas in this forum. This research was supported in parts by grants from NSF #IIS-1249822 and ONR #N00014-13-1-0153 and #N00014-10-1-0933.

References

Adams, E. (1975). *The logic of conditionals*. Dordrecht, Netherlands: D. Reidel.

Arlo-Costa, H. (2009). The logic of conditionals. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2009 ed.). `http://plato.stanford.edu/archives/spr2009/entries/logic-conditionals/`.

Balke, A., & Pearl, J. (1995). Counterfactuals and policy analysis in structural models. In P. Besnard & S. Hanks (Eds.), *Uncertainty in artificial intelligence 11* (pp. 11–18). San Francisco: Morgan Kaufmann.

Baron, R., & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Campbell, D., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research.* Chicago, IL: R. McNally and Co.

Fine, K. (1975). Review of Lewis' counterfactuals. *Mind*, *84*, 451–458.

Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundation of Science*, *3*(1), 151–182.

Gardenfors, P. (1988). Causation and the dynamics of belief. In W. Harper & B. Skyrms (Eds.), *Causation in decision, belief change and statistics II* (pp. 85–104). Kluwer Academic Publishers.

Goldszmidt, M., & Pearl, J. (1992). Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions. In B. Nebel, C. Rich, & W. Swartout (Eds.), *Proceedings of the third international conference on knowledge representation and reasoning* (pp. 661–672). San Mateo, CA: Morgan Kaufmann.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, *11*, 1–12. (Reprinted (1995) in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis* (pp. 477–490). Cambridge, MA: Cambridge University Press.)

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of themselves as interveners, not observers. *Journal of Experimental Psychology: General*, *138*, 22–38.

Halpern, J. (1998). Axiomatizing causal reasoning. In G. Cooper & S. Moral (Eds.), *Uncertainty in artificial intelligence* (pp. 202–210). San Francisco, CA: Morgan Kaufmann. (Also (2000), *Journal of Artificial Intelligence Research 12* (3), 17–37.)

Hurwicz, L. (1950). Generalization of the concept of identification. In T. Koopmans (Ed.), *Statistical inference in dynamic economic models* (pp. 245–257). New York: Wiley.

Joyce, J. (1999). *The foundatins of causal decision theory.* Cambridge, MA: Cambridge University Press.

Joyce, J. (2009). Causal reasoning and backtracking. *Philosophical Studies*, *147*, 139–154, 2010 (print).

Lewis, D. (1973). Counterfactuals and comparative probability. *Journal of Philosophical Logic*, *2*(4), 418–446. (Reprinted (1981) in W.L. Harper, R. Stalnaker, and G. Pearce (Eds.), *Ifs* (pp. 57–85). Dordrecht: D. Reidel.)

Marschak, J. (1953). Economic measurements for policy and prediction. In W. C. Hood & T. Koopmans (Eds.), *Studies in econometric method* (pp. 1–26). Wiley and Sons, Inc.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–710.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* New York: Cambridge University Press. (2nd edition, 2009.)

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2009a). *Causality: Models, reasoning, and inference* (2nd ed.). New York: Cambridge University Press.

Pearl, J. (2009b). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146. (<http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf>)

Pearl, J. (2010). *Physical and metaphysical counterfactuals* (Tech. Rep. Nos. R-359, <http://ftp.cs.ucla.edu/pub/stat_ser/r359.pdf>). CA: Department of Com-

puter Science, University of California, Los Angeles. (Forthcoming, *Review of Symbolic Logic.*)

Pearl, J. (2012a). The mediation formula: A guide to the assessment of causal pathways in nonlinear models. In C. Berzuini, P. Dawid, & L. Bernardinelli (Eds.), *Causality: Statistical perspectives and applications* (pp. 151–179). Chichester, UK: John Wiley and Sons, Ltd.

Pearl, J. (2012b). *Interpretable conditions for identifying direct and indirect effects* (Tech. Rep. Nos. R-389, <http://ftp.cs.ucla.edu/pub/stat_ser/r389.pdf>). CA: Department of Computer Science, University of California, Los Angeles.

Pearl, J. (2012c). *Do*-calculus revisited. In N. de Freitas & K. Murphy (Eds.), *Proceedings of the twenty-eighth conference on uncertainty in artificial intelligencee* (pp. 4–11). Corvallis, OR: AUAI.

Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Proceedings of the twenty-fifth conference on artificial intelligence (AAAI-11)* (pp. 247–254). Menlo Park, CA. (Available at: <http://ftp.cs.ucla.edu/pub/stat_ser/r372a.pdf>)

Pearl, J., & Verma, T. (1991). A theory of inferred causation. In J. Allen, R. Fikes, & E. Sandewall (Eds.), *Principles of knowledge representation and reasoning: Proceedings of the second international conference* (pp. 441–452). San Mateo, CA: Morgan Kaufmann.

Petersen, M., Sinisi, S., & van der Laan, M. (2006). Estimation of direct causal effects. *Epidemiology*, *17*(3), 276–284.

Ramsey, F. (1929). General propositions and causality. In F.P. Ramsey (Author) and H.A. Mellor (Ed.), *Philosophical papers* (pp. 145–153). Cambridge: Cambridge University Press.

Robins, J., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*(2), 143–155.

Shpitser, I., & Pearl, J. (2007). What counterfactuals can be tested. In *Proceedings of the twenty-third conference on uncertainty in artificial intelligence* (pp. 352–359). Vancouver, BC, Canada: AUAI Press. (Also (2008), *Journal of Machine Learning Research*, *9*, 1941–1979.)

Shpitser, I., & Pearl, J. (2009). Effects of treatment on the treated: Identification and generalization. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence* (p. http://www.cs.mcgill.ca/ uai2009/proceedings.html). Montreal, Quebec: AUAI Press.

Simon, H. (1953). Causal ordering and identifiability. In W. C. Hood & T. Koopmans (Eds.), *Studies in econometric method* (pp. 49–74). New York, NY: Wiley and Sons, Inc.

Simon, H., & Rescher, N. (1966). Cause and counterfactual. *Philosophy and Science*, *33*, 323–340.

Skyrms, B. (1980). *Causal necessity.* New Haven: Yale University Press.

Sloman, S., & Lagnado, D. (2005). Do we "do"? *Cognitive Science*, *29*, 5-39.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search.* New York: Springer-Verlag.

Stalnaker, R. (1981). Letter to David Lewis. In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 151–152). Dordrecht: D. Reidel.

Taylor, C., & Dennett, D. (2011). Who's *still* afraid of determinism? Rethinking causes and possibilities. In R. H. Kane (Ed.), *The oxford handbook of free will* (pp. 221–242). New York: Oxford University Press.

Tian, J., & Pearl, J. (2002). A general identification condition for causal effects. In *Proceedings of the eighteenth national conference on artificial intelligence* (pp. 567–573). Menlo Park, CA: AAAI Press/The MIT Press.